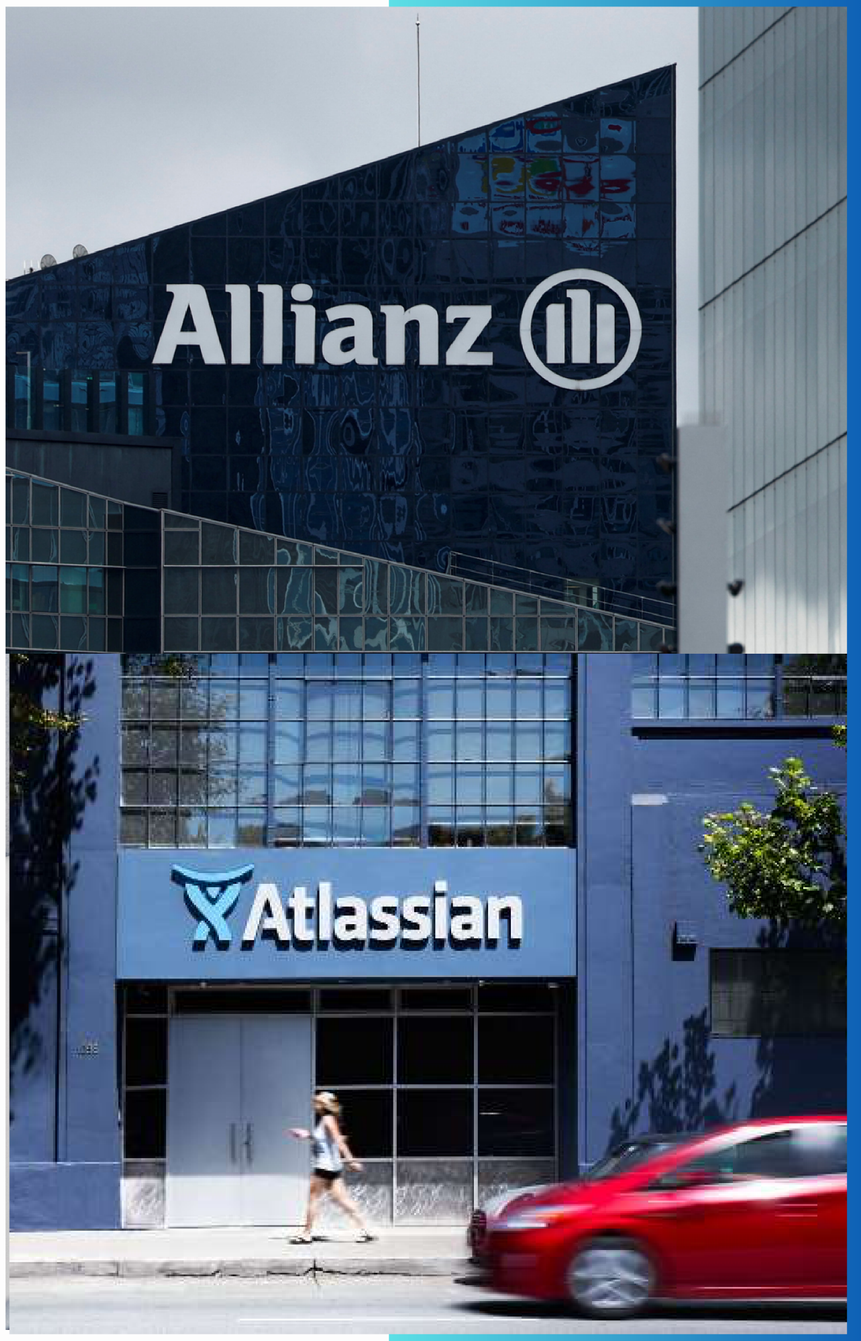
FRAUD DETECTION IN INSURANCE: DATA-DRIVEN MODEL

Atlassian x Allianz x DataSoc Datathon 2024
Group 1

Overview



Problem Statement

> $70 billion annually in Australia

~10% claims fabricated

Honest Policyholders

https://insurancecouncil.com.au/consumers/insurance-fraud/
https://www.iii.org/article/background-on-insurance-fraud

Exploratory Data Analysis

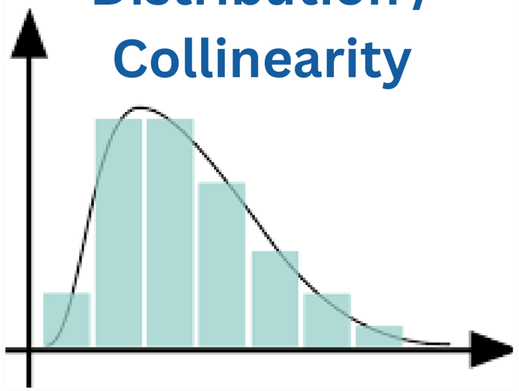
What we have found?

Unbalanced Data

Mislabeled Data

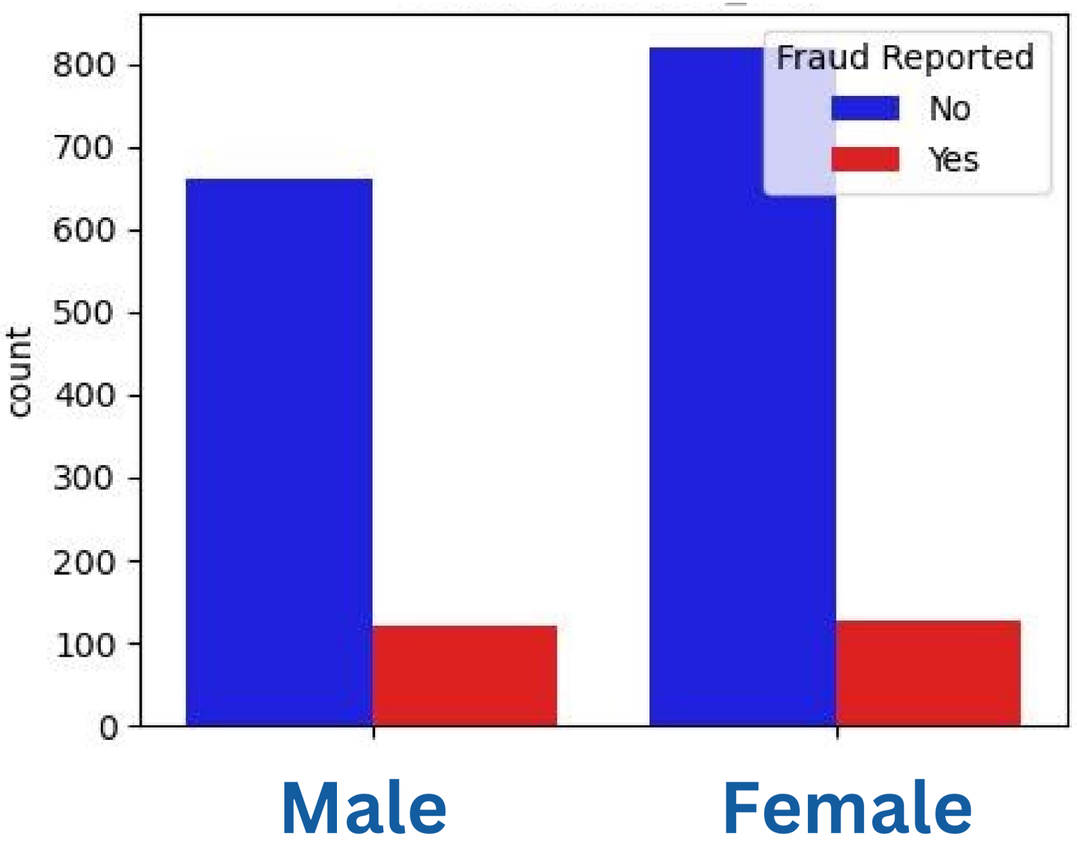
Missing Data

Skewed Distribution / Collinearity



Exploratory Data Analysis (EDA)

Gender



Sample Issues

- *Unbalanced dataset = inaccurate predictions*
- Population reflection:
  - Few commit fraud
  - Sample mirrors general behaviour

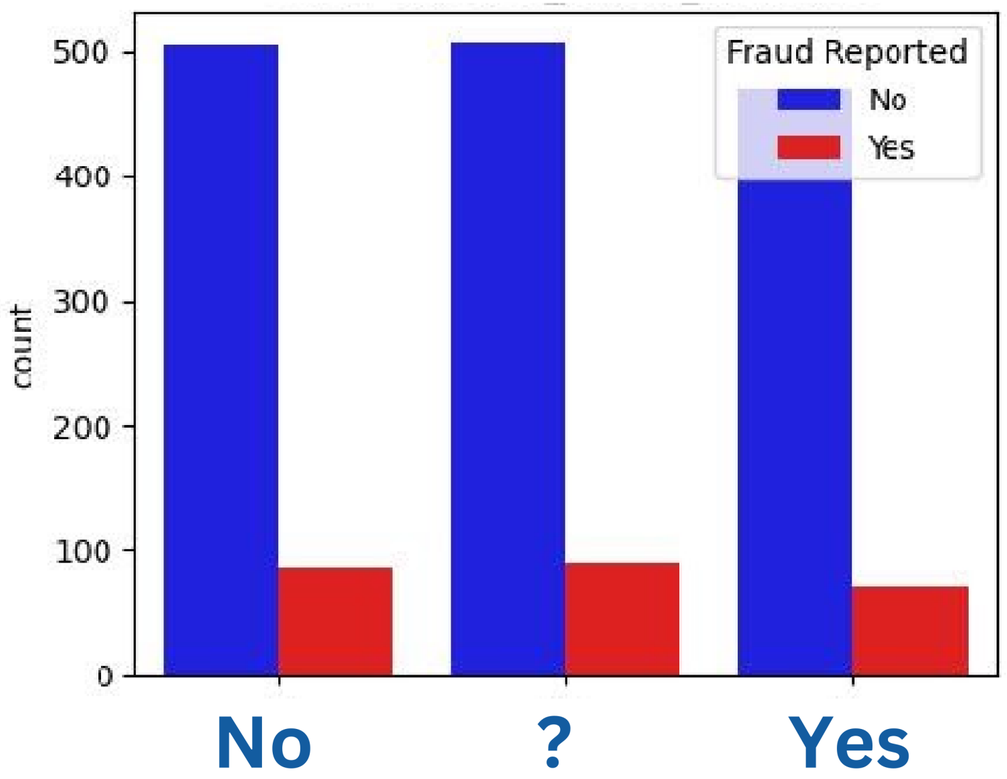
Impact on prediction

- Risks assuming no fraud in new transactions

Fraud data in all features are heavily imbalanced.

EDA - Mislabelled and Missing Data

Police Report Available



Mislabelled data:

- Accidentally missing (MAR) vs. Purposely missing (NMAR)

Impact:

- Influences model accuracy and bias

Missing Data:

- Fear of Reporting Fraud
- Coincidentally Missing Information

EDA - Mislabelled Data

Sample 1638

umbrella_limit
-1000000

How can umbrella limit be -1000000?

Feature Wrangling - Data Transformation

Education Level

```
High School     295
Masters         260
JD              258
Associate       248
MD              239
PhD             226
College         203
Name: count, dtype: int64
```

Undergraduate and below:

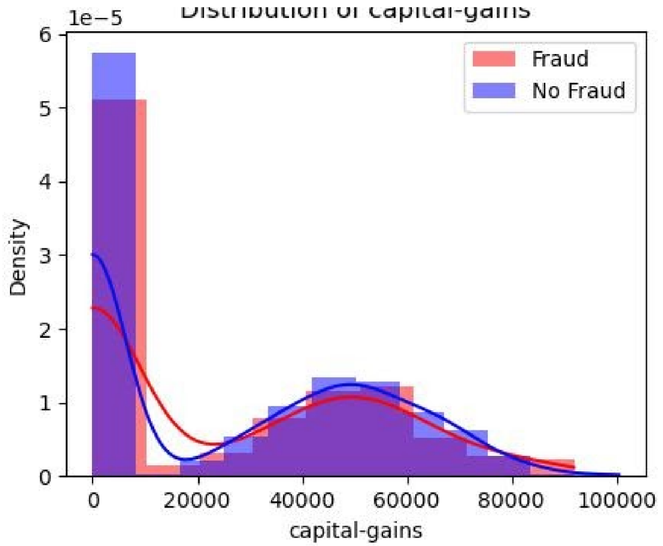
- High school
- College
- Associate

Postgraduate:

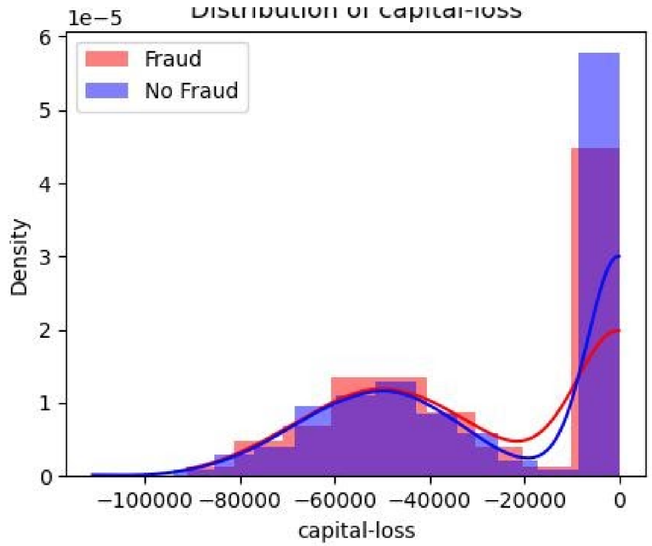
- Master’s
- JD/MD
- PhD

Feature Selection - High Collinearity

Capital Gain



Capital Loss

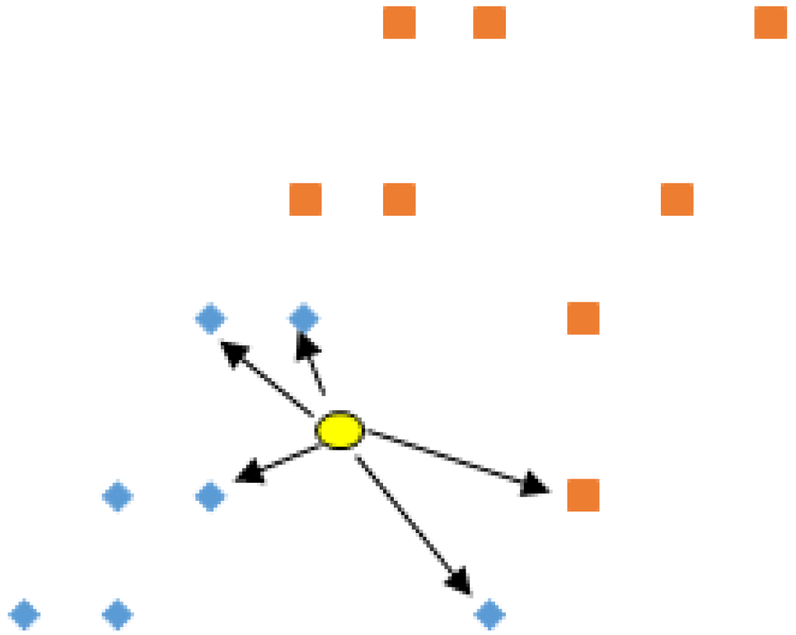


- Both metrics provide similar insights
- Same information
- Places more emphasis on this feature!

Exploratory Data Analysis (EDA)

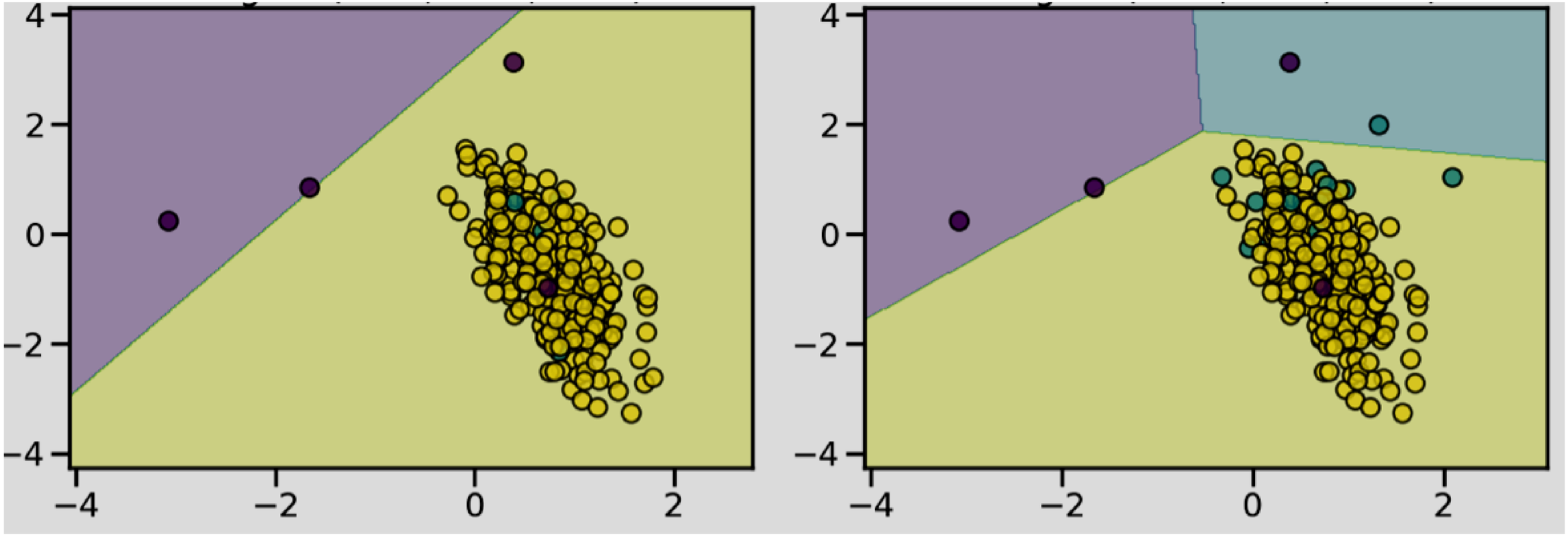
- **Scope - Does this dataset answer these questions?**
  - **Potential missing features**
    - criminal record
  - **Data considers ABC company records and not broader insurance industry**
    - Shape of Dataset
      - 1729
    - Oversampling
      - SMOTE-N

Exploratory Data Analysis (EDA)



- **SMOTE-n**
  - Randomly selects qualitative variables from fraud labels
  - finds k-nearest neighbours
  - synthetically makes new datapoint on line
- **why ?**
  - Doesn’t duplicate samples exactly

Exploratory Data Analysis (EDA)



Before **After**

ref : sklearn - imbalanced dataset

Model Selection

Generalized Linear Models (GLMs) and Additive Models

- Interpretability
- Scalability

- Strict Assumptions
- Simplicity

Random Forest and Boostings

- Nonlinear
- Interpretability
- No Assumptions

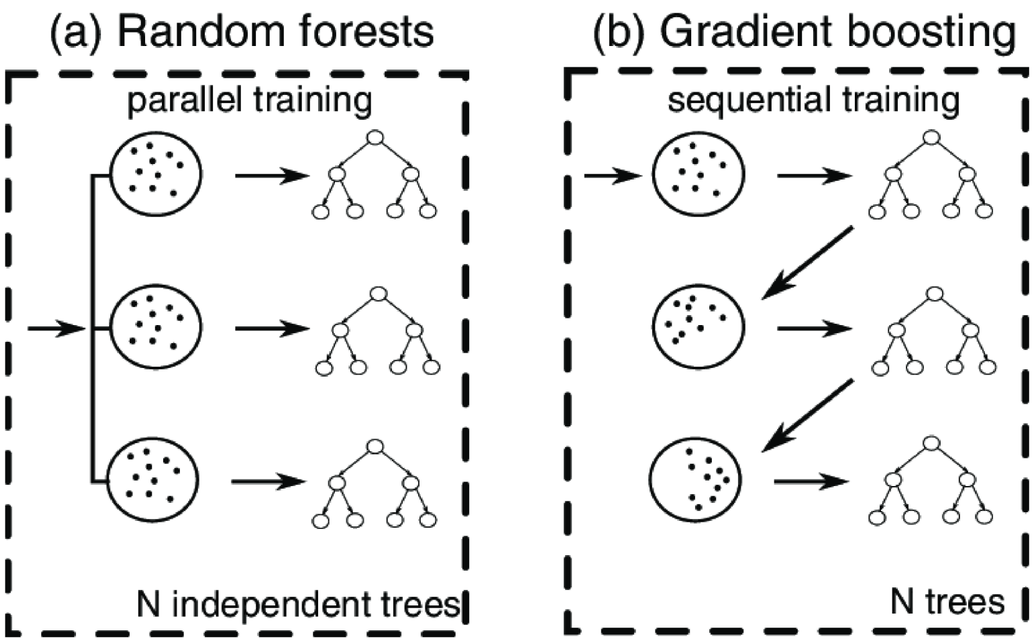
- Overfitting

Neural Networks

- Flexibility
- High Accuracy

- Black box Nature
- Large Dataset

Modelling and Tuning



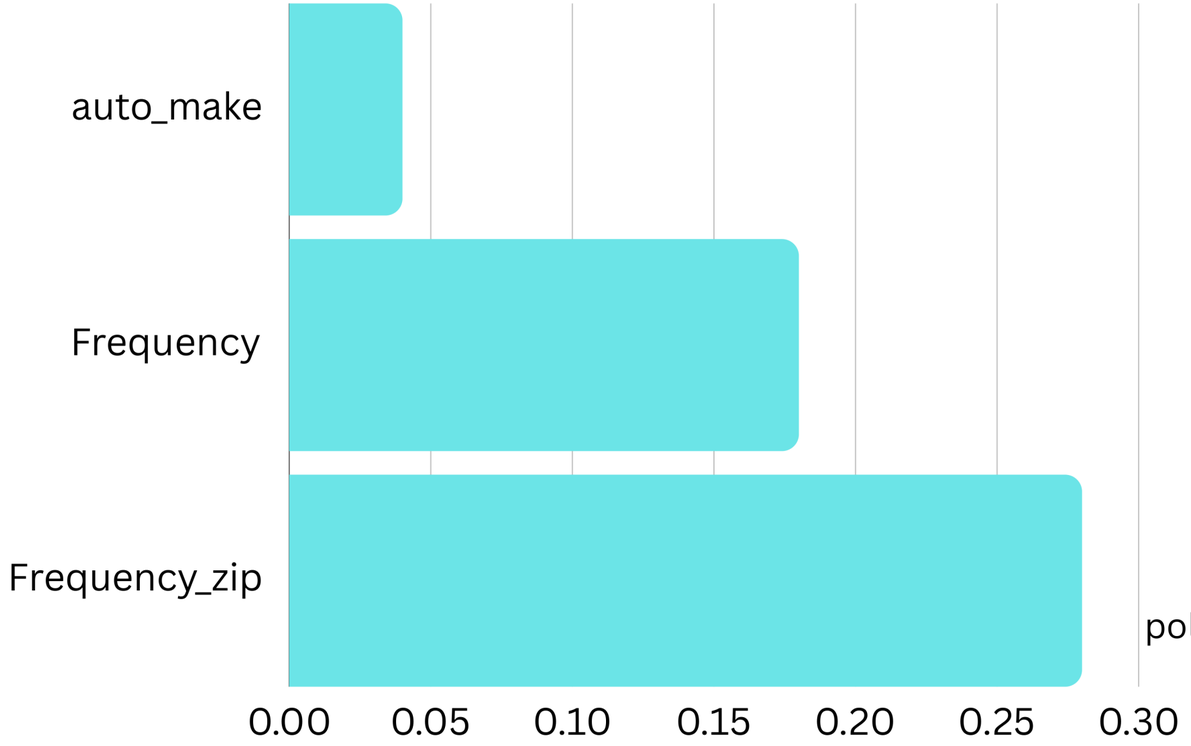
- **Random forest, XGBoost, lightGBM and ensemble model were used.**
- **To build optimal machine learning models, hyperparameter tuning is important and it was conducted using optuna.**
- **Since it is probabilistic model, R2 score was used to evaluate models in this research.**

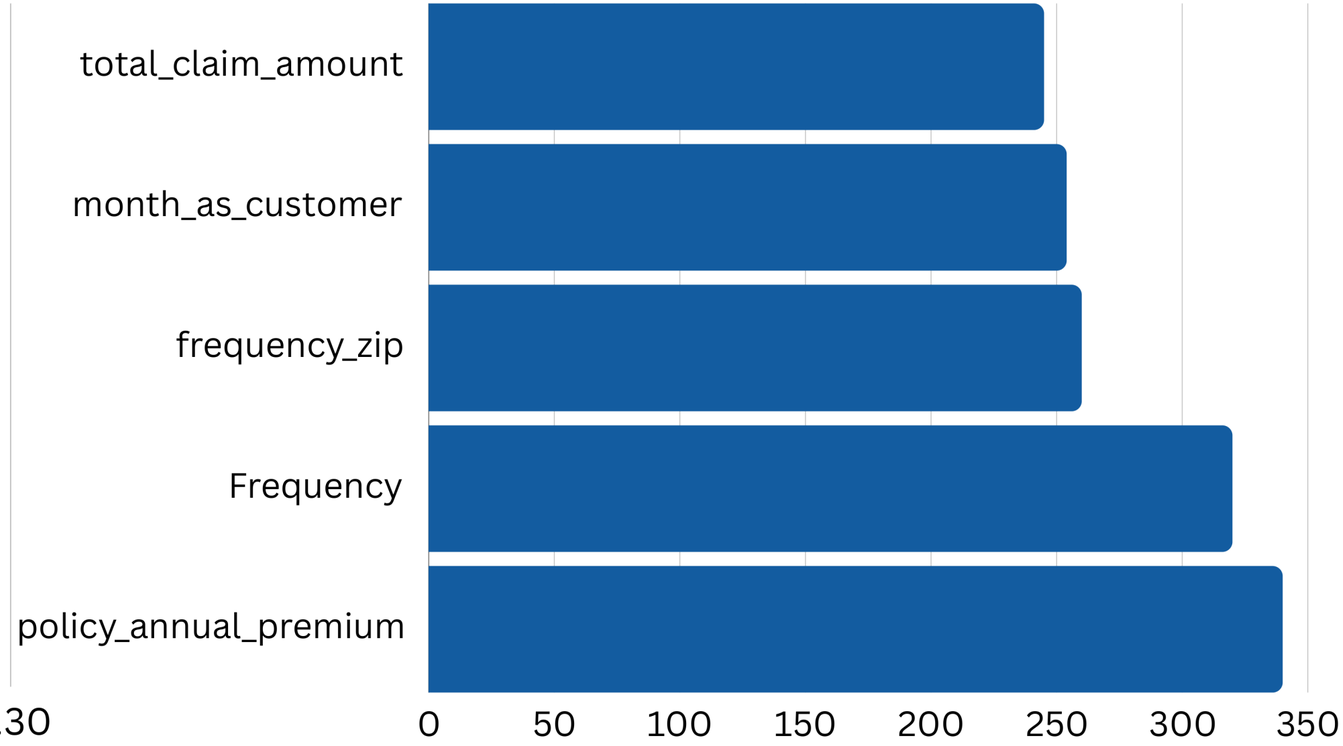
Model Performance

	Without Ethical Considerations	With Ethical Considerations
Random Forest	0.92382	0.89516
XgBoost	0.89884	0.88923
LightGBM	0.91138	0.88118
Ensemble	0.91842	0.88603

- **Random Forest achieved the highest R2 scores among models, with and without ethical considerations.**
- **Random Forest and Ensemble methods required significantly more computation time compared to XGBoost and LightGBM.**
- **The computational complexity of Random Forest becomes a concern, especially with large original datasets.**

Feature Importances





- **In both Random Forest and XGBoost models, the frequency of zip codes and policy numbers emerged as the most important features.**

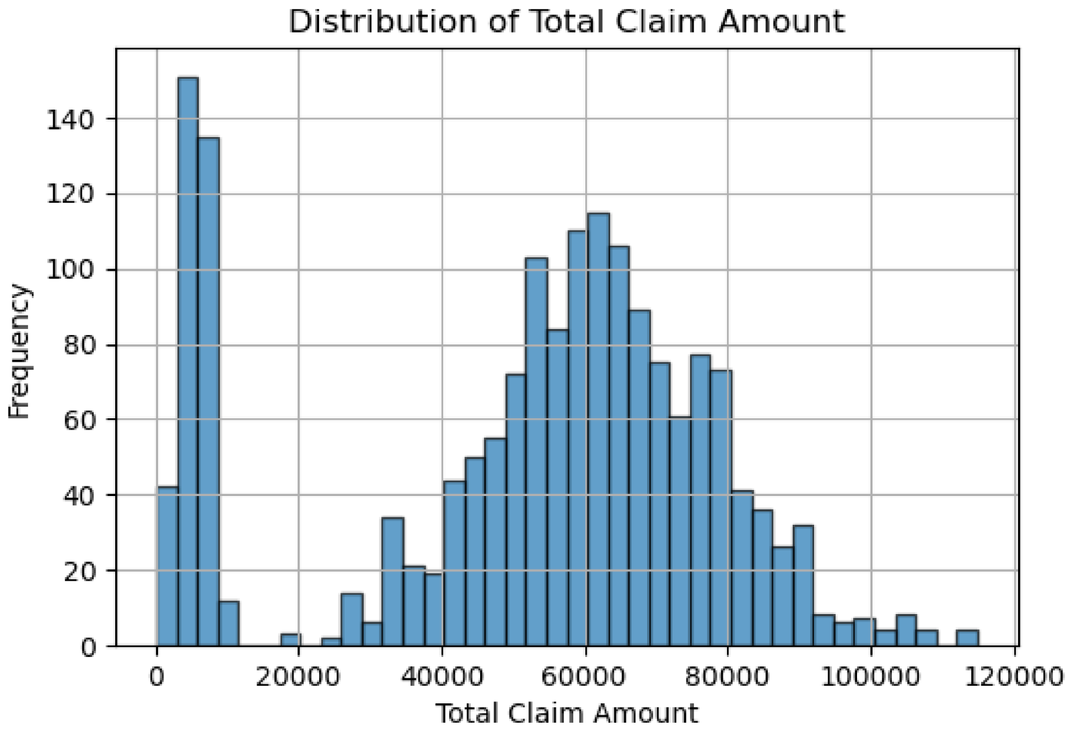
- **In lightGBM, adding to those frequencies, total claim amount, month_as_customers and policy annual premium were also important.**

Business Solution

- Two Decisions for ML model outcomes

Balance between Cost and Potential Saving

- Assume the average total investigating cost per review case: AUD5,000



Proportion of a fraud case

- overall proportion of Fraud: 14.3%
- Total claim amount cost below 20k: 4.66%
- Total claim amount cost above 20k: 16.67%

Recommendation for Below 20k

- There are only 16 fraud out of 343 cases
- Frequency = 1
- Average amount = $5,145

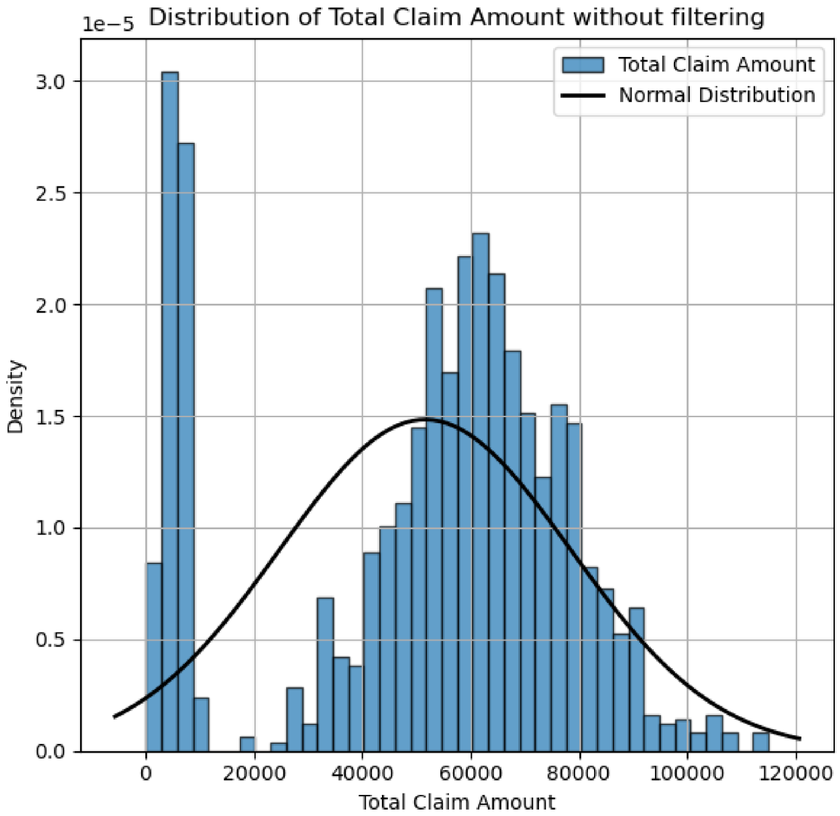
Recommendation for Above 20k

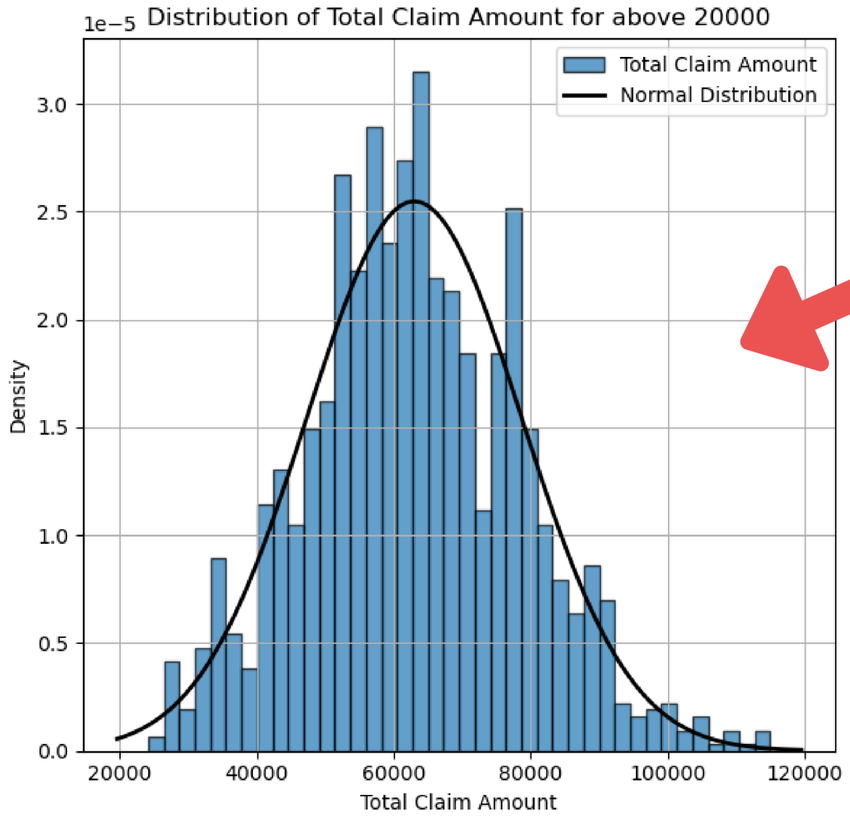
As our ML models applied SmoteN for oversampling, we need to use Bayes' Theorem for adjustment. Prior=0.143

- Bayes' Theorem:

$$P(H|E) = \frac{P(E|H) \cdot P(H)}{P(E|H) \cdot P(H) + P(E|\neg H) \cdot P(\neg H)}$$

Recommendation for Above 20k





Mean: 63,000

Recommendation for Above 20k

Example

Expected total claim amount (Potential Savings)

= $63,000 x 0.5

= $31,500 >>> Investgating cost

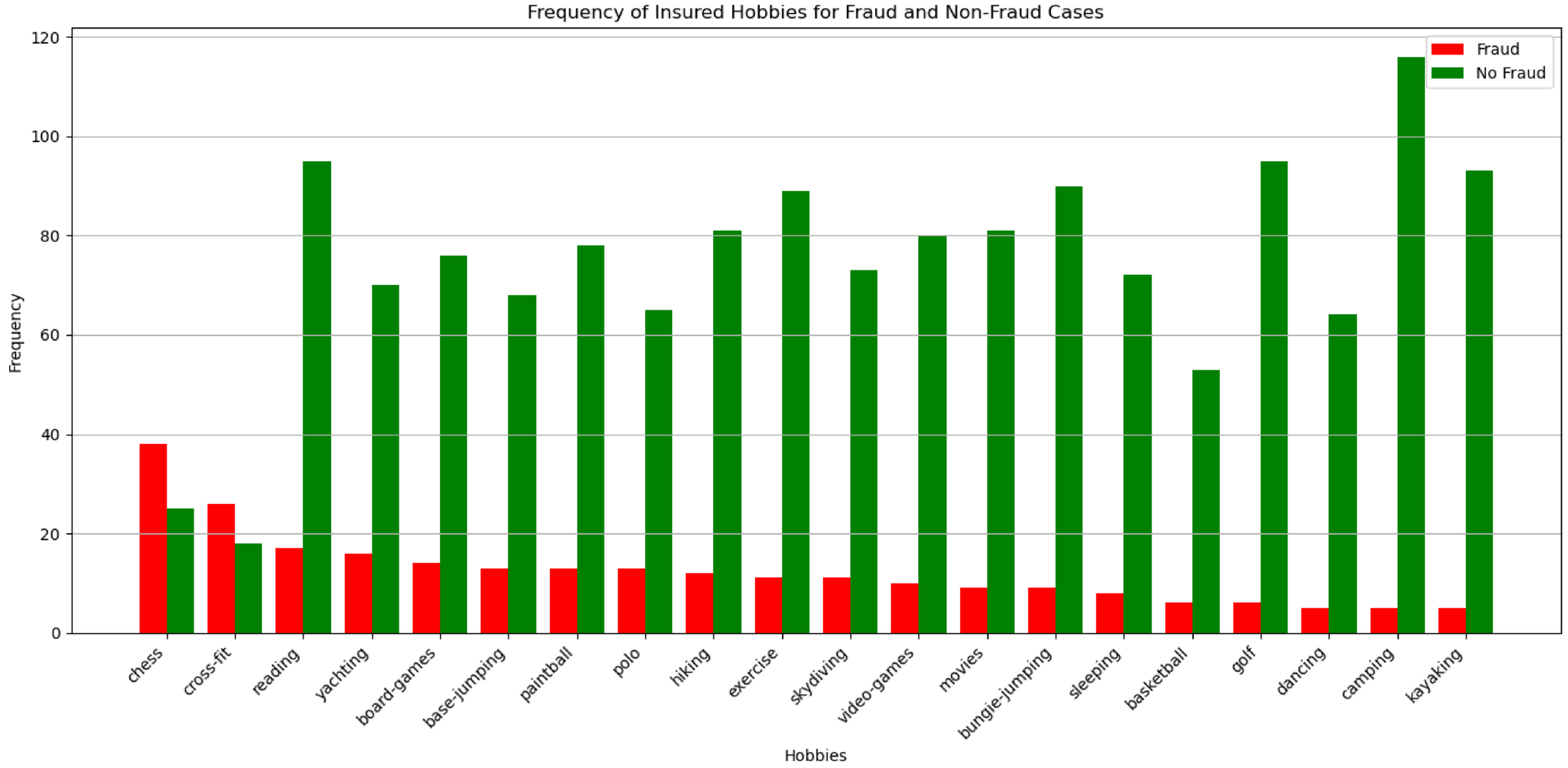
Worth for Reviewing predict probability >50% cases.

Review

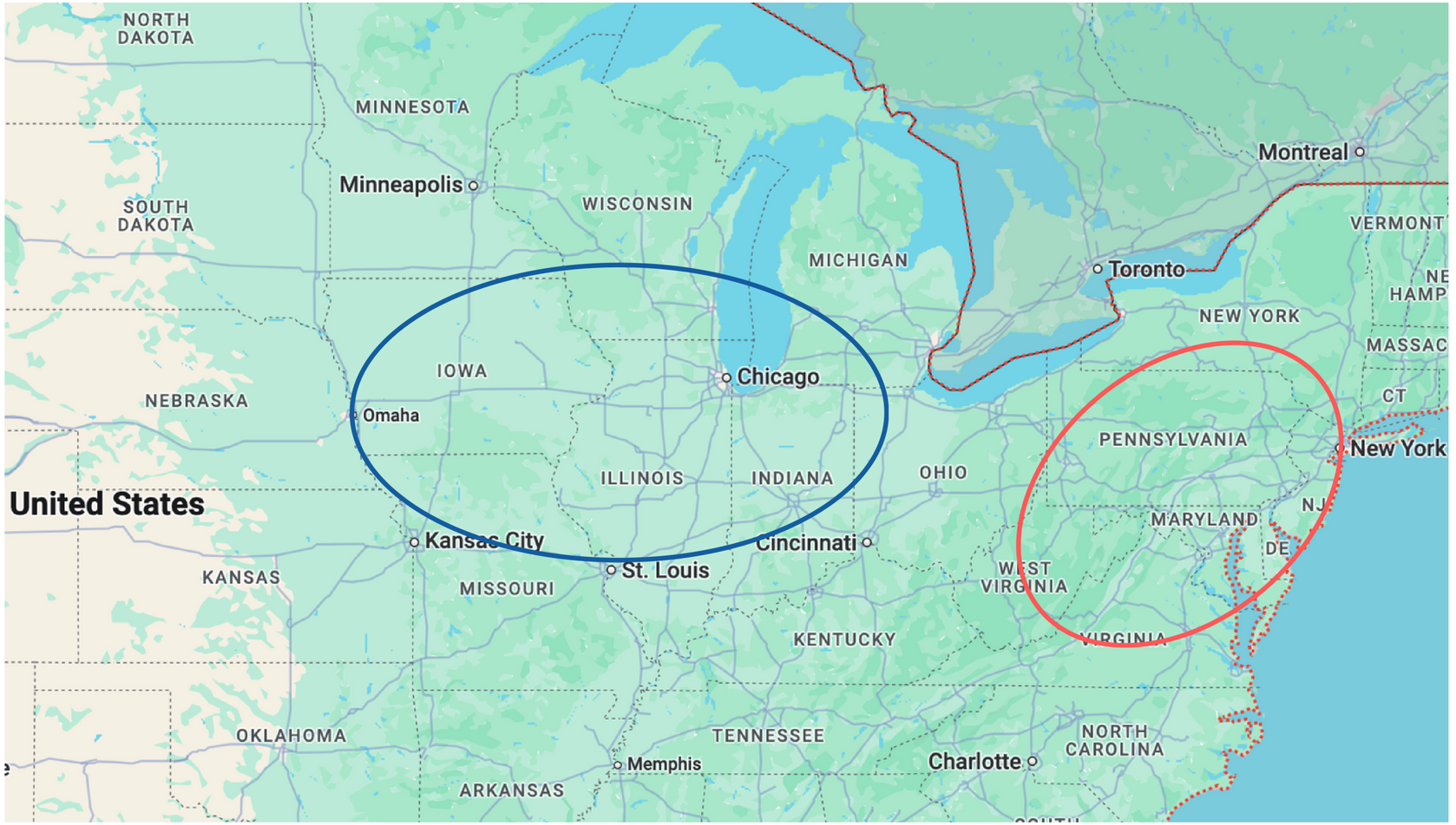
Ethics - Gender (Legal)

- In Europe, gender is prohibited as a predictor
- Not significant in our models.

Ethics - Hobbies (PR matter)



Limitation - Geographic Limitation



Thank you!

Q&A time!